Introduction to Probability[1]

(2013 - I)

Salvador Ruiz Correa

Centro de Investigación en Matemáticas (CIMAT)

[1] These slides are *adapted* from those that accompany the book *Bayesian Reasoning and Machine Learning*. The book and demos can be downloaded from `www.cs.ucl.ac.uk/staff/D.Barber/brml`. We acknowledge David Barber for providing the original slides.

Probability space $(\Omega, \mathscr{F}, P)$

A probability space consists of an arbitrary set Ω called sample space, a σ-algebra $\mathscr{F}$, and a probability measure P.

σ-algebra

A collection $\mathscr{F}$ of subsets of Ω is a σ-algebra if the following conditions are true:

1. $\Omega \in \mathscr{F}$
2. If $A \in \mathcal{F}$, then $A^c \in \mathscr{F}$.
3. If $A_1, A_2, \cdots \in \mathscr{F}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathscr{F}$

Namely, a σ-algebra is a nonempty collection of subsets of Ω that is closed under complement and infinite unions. These properties guarantee that the collection is closed under union, intersection, complement, difference, symmetric difference, and so on. The elements of $\mathscr{F}$ are called *events*.

Example: Borel algebras in $\mathbb{R}$.

Borel algebra

- Definition
$$\mathscr{B}(\mathscr{R}) = \sigma\{(a, b) \subseteq \mathbb{R} : a \leq b\}.$$
- The elements of $\mathscr{B}(\mathbb{R})$ are calle Borel sets.
- For any real numbers $a \leq b$, the intervals $[a, b]$, $(-\infty, b)$, $[a, b)$, $(a, b]$ and $\{a\}$, are all elements of $\mathscr{B}(\mathbb{R})$.

Probability measure

Let $(\Omega, \mathscr{F})$ a measurable space. A probability measure is a function $P : \mathscr{F} \to [0, 1]$ that satisfies:

1. $P(\Omega) = 1$.
2. If $P(A) \geq 0$, for all $A^c \in \mathscr{F}$.
3. If $A_1, A_2, \dots \in \mathscr{F}$, are jointly disjoint, that is, $A_m \cap A_n = \emptyset$ for $m \neq n$, then $P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$.

Random variable

A real random variable is a function $X : \Omega \to \mathbb{R}$ such that for any Borel set B, the set $X^{-1}B$ is an element of $\mathscr{F}$.

Rules of probability discrete random variables

$p(x = \mathsf{x})$: the probability of variable x being in state x.

$$p(x = \mathsf{x}) = \begin{cases} 1 & \text{we are certain } x \text{ is in state x} \\ 0 & \text{we are certain } x \text{ is not in state x} \end{cases}$$

Values between 0 and 1 represent the degree of certainty of state occupancy.

---

domain

$\mathrm{dom}(x)$ denotes the states x can take. For example, $\mathrm{dom}(c) = \{\text{heads}, \text{tails}\}$. When summing over a variable $\sum_x f(x)$, the interpretation is that all states of x are included, *i.e.* $\sum_x f(x) \equiv \sum_{\mathsf{s} \in \mathrm{dom}(x)} f(x = \mathsf{s})$.

---

distribution

Given a variable, x, its domain $\mathrm{dom}(x)$ and a full specification of the probability values for each of the variable states, $p(x)$, we have a distribution for x.

---

normalisation

The summation of the probability over all the states is 1:

$$\sum_{x \in \mathrm{dom}(x)} p(x = \mathsf{x}) = 1$$

We will usually more conveniently write $\sum_x p(x) = 1$.

Operations

AND

Use the shorthand $p(x, y) \equiv p(x \cap y)$ for $p(x$ and $y)$. Note that $p(y, x) = p(x, y)$.

Marginalisation

Given a joint distr. $p(x, y)$ the marginal distr. of x is defined by

$$p(x) = \sum_y p(x, y)$$

More generally,

$$p(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = \sum_{x_i} p(x_1, \ldots, x_n)$$

Conditional Probability and Bayes' Rule

The probability of event x conditioned on knowing event y (or more shortly, the probability of x given y) is defined as

$$p(x|y) \equiv \frac{p(x,y)}{p(y)} = \frac{p(y|x)p(x)}{p(y)} \quad \text{(Bayes' rule)}$$

Throwing darts

$$\begin{aligned} p(\text{region } 5|\text{not region } 20) &= \frac{p(\text{region } 5, \text{not region } 20)}{p(\text{not region } 20)} \\ &= \frac{p(\text{region } 5)}{p(\text{not region } 20)} = \frac{1/20}{19/20} = \frac{1}{19} \end{aligned}$$

Interpretation

$p(A = \mathrm{a}|B = \mathrm{b})$ should not be interpreted as 'Given the event $B = \mathrm{b}$ has occurred, $p(A = \mathrm{a}|B = \mathrm{b})$ is the probability of the event $A = \mathrm{a}$ occurring'. The correct interpretation should be '$p(A = \mathrm{a}|B = \mathrm{b})$ is the probability of A being in state a under the constraint that B is in state b'.

Probability tables

The a priori probability that a randomly selected Great British person would live in England, Scotland or Wales, is 0.88, 0.08 and 0.04 respectively.

We can write this as a vector (or probability table) :

$$\left(\begin{array}{c} p(Cnt = \mathsf{E}) \\ p(Cnt = \mathsf{S}) \\ p(Cnt = \mathsf{W}) \end{array} \right) = \left(\begin{array}{c} 0.88 \\ 0.08 \\ 0.04 \end{array} \right)$$

whose component values sum to 1.

The ordering of the components in this vector is arbitrary, as long as it is consistently applied.

Probability tables

We assume that only three Mother Tongue languages exist : English (Eng), Scottish (Scot) and Welsh (Wel), with conditional probabilities given the country of residence, England (E), Scotland (S) and Wales (W). Using the state ordering:

$$MT = [\mathsf{Eng}, \mathsf{Scot}, \mathsf{Wel}]; \qquad Cnt = [\mathsf{E}, \mathsf{S}, \mathsf{W}]$$

we write a (fictitious) conditional probability table

$$p(MT|Cnt) = \begin{pmatrix} 0.95 & 0.7 & 0.6 \\ 0.04 & 0.3 & 0.0 \\ 0.01 & 0.0 & 0.4 \end{pmatrix}$$

Probability tables

The distribution $p(Cnt, MT) = p(MT|Cnt)p(Cnt)$ can be written as a 3×3 matrix with (say) rows indexed by country and columns indexed by Mother Tongue:

$$\begin{pmatrix} 0.95 \times 0.88 & 0.7 \times 0.08 & 0.6 \times 0.04 \\ 0.04 \times 0.88 & 0.3 \times 0.08 & 0.0 \times 0.04 \\ 0.01 \times 0.88 & 0.0 \times 0.08 & 0.4 \times 0.04 \end{pmatrix} = \begin{pmatrix} 0.836 & 0.056 & 0.024 \\ 0.0352 & 0.024 & 0 \\ 0.0088 & 0 & 0.016 \end{pmatrix}$$

By summing a column, we have the marginal

$$p(Cnt) = \begin{pmatrix} 0.88 \\ 0.08 \\ 0.04 \end{pmatrix}$$

Summing the rows gives the marginal

$$p(MT) = \begin{pmatrix} 0.916 \\ 0.0592 \\ 0.0248 \end{pmatrix}$$

Probability tables

Large numbers of variables

For joint distributions over a larger number of variables, $x_i, i = 1, \ldots, D$, with each variable x_i taking K_i states, the table describing the joint distribution is an array with $\prod_{i=1}^{D} K_i$ entries.

Explicitly storing tables therefore requires space exponential in the number of variables, which rapidly becomes impractical for a large number of variables.

Indexing

A probability distribution assigns a value to each of the joint states of the variables. For this reason, $p(T, J, R, S)$ is considered equivalent to $p(J, S, R, T)$ (or any such reordering of the variables), since in each case the joint setting of the variables is simply a different index to the same probability.

One should be careful not to confuse the use of this indexing type notation with functions $f(x, y)$ which are in general dependent on the variable order.

Inspector Clouseau

Inspector Clouseau arrives at the scene of a crime. The Butler (B) and Maid (M) are his main suspects. The inspector has a prior belief of 0.6 that the Butler is the murderer, and a prior belief of 0.2 that the Maid is the murderer. These probabilities are independent in the sense that $p(B, M) = p(B)p(M)$. (It is possible that both the Butler and the Maid murdered the victim or neither). The inspector's *prior* criminal knowledge can be formulated mathematically as follows:

$$\mathrm{dom}(B) = \mathrm{dom}(M) = \{\text{murderer}, \text{not murderer}\}$$

$$\mathrm{dom}(K) = \{\text{knife used}, \text{knife not used}\}$$

$$p(B = \text{murderer}) = 0.6, \qquad p(M = \text{murderer}) = 0.2$$

$$\begin{aligned}
p(\text{knife used}|B = \text{not murderer}, \quad & M = \text{not murderer}) &= 0.3 \\
p(\text{knife used}|B = \text{not murderer}, \quad & M = \text{murderer}) &= 0.2 \\
p(\text{knife used}|B = \text{murderer}, \quad & M = \text{not murderer}) &= 0.6 \\
p(\text{knife used}|B = \text{murderer}, \quad & M = \text{murderer}) &= 0.1
\end{aligned}$$

The victim lies dead in the room and the inspector quickly finds the murder weapon, a Knife (K). What is the probability that the Butler is the murderer? (Remember that it might be that neither is the murderer).

Inspector Clouseau

Using b for the two states of B and m for the two states of M,

$$p(B|K) = \sum_m p(B,m|K) = \sum_m \frac{p(B,m,K)}{p(K)} = \frac{p(B)\sum_m p(K|B,m)p(m)}{\sum_b p(b)\sum_m p(K|b,m)p(m)}$$

Plugging in the values we have

$$\begin{aligned} p(B = \text{murderer}|\text{knife used}) &= \frac{\frac{6}{10}\left(\frac{2}{10}\times\frac{1}{10}+\frac{8}{10}\times\frac{6}{10}\right)}{\frac{6}{10}\left(\frac{2}{10}\times\frac{1}{10}+\frac{8}{10}\times\frac{6}{10}\right)+\frac{4}{10}\left(\frac{2}{10}\times\frac{2}{10}+\frac{8}{10}\times\frac{3}{10}\right)} \\ &= \frac{300}{412} \approx 0.73 \end{aligned}$$

Hence knowing that the knife was the murder weapon strengthens our belief that the butler did it.

Exercise: compute the probability that the Butler and not the Maid is the murderer.

Inspector Clouseau

The role of $p(\text{knife used})$ in the Inspector Clouseau example can cause some confusion. In the above,

$$p(\text{knife used}) = \sum_b p(b) \sum_m p(\text{knife used}|b, m)p(m)$$

is computed to be 0.456. But surely, $p(\text{knife used}) = 1$, since this is given in the question!

Note that the quantity $p(\text{knife used})$ relates to the *prior* probability the model assigns to the knife being used (in the absence of any other information). If we know that the knife is used, then the *posterior* is

$$p(\text{knife used}|\text{knife used}) = \frac{p(\text{knife used}, \text{knife used})}{p(\text{knife used})} = \frac{p(\text{knife used})}{p(\text{knife used})} = 1$$

which, naturally, must be the case.

Independence

Variables x and y are independent if knowing one event gives no extra information about the other event. Mathematically, this is expressed by

$$p(x, y) = p(x)p(y)$$

Independence of x and y is equivalent to

$$p(x|y) = p(x) \Leftrightarrow p(y|x) = p(y)$$

If $p(x|y) = p(x)$ for all states of x and y, then the variables x and y are said to be independent. We write then $x \perp\!\!\!\perp y$.

---

interpretation

Note that $x \perp\!\!\!\perp y$ doesn't mean that, given y, we have no information about x. It means the only information we have about x is contained in $p(x)$.

---

factorisation

If

$$p(x, y) = kf(x)g(y)$$

for some constant k, and positive functions $f(\cdot)$ and $g(\cdot)$ then x and y are independent.

Conditional Independence

$$\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \mathcal{Z}$$

denotes that the two sets of variables $\mathcal{X}$ and $\mathcal{Y}$ are independent of each other given the state of the set of variables $\mathcal{Z}$. This means that

$$p(\mathcal{X}, \mathcal{Y} | \mathcal{Z}) = p(\mathcal{X} | \mathcal{Z}) p(\mathcal{Y} | \mathcal{Z}) \text{ and } p(\mathcal{X} | \mathcal{Y}, \mathcal{Z}) = p(\mathcal{X} | \mathcal{Z})$$

for all states of $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$. In case the conditioning set is empty we may also write $\mathcal{X} \perp\!\!\!\perp \mathcal{Y}$ for $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \emptyset$, in which case $\mathcal{X}$ is (unconditionally) independent of $\mathcal{Y}$.

Conditional independence does not imply marginal independence

$$p(x, y) = \sum_z \underbrace{p(x|z)p(y|z)}_{\text{cond. indep.}} p(z) \neq \underbrace{\sum_z p(x|z)p(z)}_{p(x)} \underbrace{\sum_z p(y|z)p(z)}_{p(y)}$$

Conditional dependence

If $\mathcal{X}$ and $\mathcal{Y}$ are not conditionally independent, they are conditionally dependent. This is written

$$\mathcal{X} \top\!\!\!\top \mathcal{Y} | \mathcal{Z}$$

Conditional Independence example

Based on a survey of households in which the husband and wife each own a car, it is found that:

$$\text{wife's car type} \perp\!\!\!\perp \text{husband's car type} | \text{family income}$$

There are 4 car types, the first two being 'cheap' and the last two being 'expensive'. Using w for the wife's car type and h for the husband's:

$$p(w|inc = \text{low}) = \begin{pmatrix} 0.7 \\ 0.3 \\ 0 \\ 0 \end{pmatrix}, \quad p(w|inc = \text{high}) = \begin{pmatrix} 0.2 \\ 0.1 \\ 0.4 \\ 0.3 \end{pmatrix}$$

$$p(h|inc = \text{low}) = \begin{pmatrix} 0.2 \\ 0.8 \\ 0 \\ 0 \end{pmatrix}, \quad p(h|inc = \text{high}) = \begin{pmatrix} 0 \\ 0 \\ 0.3 \\ 0.7 \end{pmatrix}$$

$$p(inc = \text{low}) = 0.8$$

Conditional Independence example

Then the marginal distribution $p(w,h)$ is

$$p(w,h) = \sum_{inc} p(w|inc)p(h|inc)p(inc)$$

giving

$$p(w,h) = \begin{pmatrix} 0.126 & 0.504 & 0.006 & 0.014 \\ 0.054 & 0.216 & 0.003 & 0.007 \\ 0 & 0 & 0.012 & 0.028 \\ 0 & 0 & 0.009 & 0.021 \end{pmatrix}$$

From this we can find the marginals and calculate

$$p(w)p(h) = \begin{pmatrix} 0.117 & 0.468 & 0.0195 & 0.0455 \\ 0.0504 & 0.2016 & 0.0084 & 0.0196 \\ 0.0072 & 0.0288 & 0.0012 & 0.0028 \\ 0.0054 & 0.0216 & 0.0009 & 0.0021 \end{pmatrix}$$

This shows that whilst $w \perp\!\!\!\perp h | inc$, it is not true that $w \perp\!\!\!\perp h$. For example, even if we don't know the family income, if we know that the husband has a cheap car then his wife must also have a cheap car – these variables are therefore dependent.

Scientific Inference

Much of science deals with problems of the form : tell me something about the variable θ given that I have observed data $\mathcal{D}$ and have some knowledge of the underlying data generating mechanism. Our interest is then the quantity

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int_\theta p(\mathcal{D}|\theta)p(\theta)}$$

This shows how from a forward or *generative model* $p(\mathcal{D}|\theta)$ of the dataset, and coupled with a *prior* belief $p(\theta)$ about which variable values are appropriate, we can infer the *posterior* distribution $p(\theta|\mathcal{D})$ of the variable in light of the observed data.

Generative models in science

This use of a generative model sits well with physical models of the world which typically postulate how to generate observed phenomena, assuming we know the model. For example, one might postulate how to generate a time-series of displacements for a swinging pendulum but with unknown mass, length and damping constant. Using this generative model, and given only the displacements, we could infer the unknown physical properties of the pendulum, such as its mass, length and friction damping constant.

Prior, Likelihood and Posterior

For data $\mathcal{D}$ and variable θ, Bayes' rule tells us how to update our prior beliefs about the variable θ in light of the data to a posterior belief:

$$\underbrace{p(\theta|\mathcal{D})}_{\text{posterior}} = \frac{\overbrace{p(\mathcal{D}|\theta)}^{}\!\!\!\!\!\!\underbrace{\phantom{p(\mathcal{D}|\theta)}}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}}{\underbrace{p(\mathcal{D})}_{\text{evidence}}}$$

The evidence is also called the marginal likelihood.

The term likelihood is used for the probability that a model generates observed data.

Prior, Likelihood and Posterior

More fully, if we condition on the model M, we have

$$p(\theta|\mathcal{D}, M) = \frac{p(\mathcal{D}|\theta, M)p(\theta|M)}{p(\mathcal{D}|M)}$$

where we see the role of the likelihood $p(\mathcal{D}|\theta, M)$ and marginal likelihood $p(\mathcal{D}|M)$. The marginal likelihood is also called the model likelihood.

The MAP assignment

The Most probable A Posteriori (MAP) setting is that which maximises the posterior,

$$\theta_* = \underset{\theta}{\operatorname{argmax}}\, p(\theta|\mathcal{D}, M) = \underset{\theta}{\operatorname{argmax}}\, p(\theta, \mathcal{D}|M)$$

The Max Likelihood assignment

When $p(\theta|M) = \text{const.}$,

$$\theta_* = \underset{\theta}{\operatorname{argmax}}\, p(\theta, \mathcal{D}|M) = \underset{\theta}{\operatorname{argmax}}\, p(\mathcal{D}|\theta, M)$$

Homework

Problems 1.6, 1.7, 1.9, 1.18, and 1.19.